# 5 Análisis de resultados

## 5.1 Introducción

En este capítulo se describe el procedimiento empleado para determinar la validación de las ideas expuestas anteriormente. Primero se describe el tipo de evaluación a efectuar y las métricas seleccionadas: precisión, especificidad, ***F*** y exactitud. Después, se presenta la metodología experimental para determinar el tamaño apropiado de la muestra que permite probar la hipótesis planteada con el prototipo inicial.

Posteriormente se describe cronológicamente el proceso de obtención de resultados para poder apreciar los criterios que guiaron la mejora paulatina del sistema. Así, en tercer lugar se describe el proceso y resultados obtenidos de las corridas del sistema utilizando archivos de texto etiquetado y diccionarios específicos (ad hoc) de sinónimos y escenarios. En cuarto lugar se presentan los resultados de las corridas en archivos seleccionados y verificados manualmente. Después se presenta el análisis para determinar el tamaño apropiado para la ventana de búsqueda hacia atrás para reducir el tiempo de respuesta. Finalmente se comentan los resultados de los experimentos realizados con archivos de texto libre que permiten obtener las conclusiones presentadas en el siguiente capítulo.

## 5.2 Métricas seleccionadas

La evaluación a realizar es ***intrínseca o de categorización*** porque juzga la calidad o efectividad del sistema en la asignación automática de la expresión nominal, a la categoría de correferencia, anáfora indirecta o referencia, comparando los resultados con la asignación o verificación manual de un "experto" humano. Para la medición de resultados se seleccionó como métricas primarias la ***precisión*** y *la* ***especificidad*** (del Inglés precision y recall respectivamente). Entendiendo como ***precisión*** la habilidad del sistema de identificar ***sólo*** *los elementos relevantes*

(o pertenecientes a la categoría), y como ***especificidad*** la de identificar ***todos*** *los elementos relevantes* [Salton, 1989]. Sin embargo, estas métricas, por si solas, no permiten apreciar el comportamiento y rendimiento total del sistema; esto hace necesario utilizar, como métrica adicional para apreciar el comportamiento global del sistema, la exactitud en el acuerdo (del Inglés agreement) o coincidencia entre el sistema y el experto. La ***exactitud***, para los fines de este trabajo, se entiende como la capacidad del sistema de identificar los elementos, tanto relevantes como no relevantes, "*de acuerdo con el experto*". Para explicar mejor estás métricas se utilizará una tabla para decisiones de clasificación como se muestra en la tabla 19.

| | | **EXPERTO** | | |
|---|---|---|---|---|
| | | **Si** | **No** | |
| **SISTEMA** | **Si** | a | b | k = a + b |
| | **No** | c | d | m = c + d |
| | | r = a + c | s = b + d | n = a + b + c + d |

**Tabla 19 Contingencias para decisiones de clasificación**

A continuación, de acuerdo a la tabla 19, se establecen las fórmulas para las métricas propuestas.

$$precisión = \frac{a}{k} \qquad especificidad = \frac{a}{r} \qquad exactitud = \frac{a+d}{n}$$

Donde:

a = la proporción de elementos **asignados** a la categoría por el sistema **y que si son** miembros de esa categoría de acuerdo con el "experto"

b = la proporción de elementos **asignados** a la categoría por el sistema **y que no son** miembros de esa categoría de acuerdo con el "experto"

c = la proporción de elementos **no asignados** a la categoría por el sistema **y que si son** miembros de esa categoría de acuerdo con el "experto"

d = la proporción de elementos **no asignados** a la categoría por el sistema **y que no son** miembros de esa categoría de acuerdo con el "experto"

k = suma de todos los elementos **asignados** a la categoría **por el sistema**

m = suma de todos los elementos **no asignados** a la categoría **por el sistema**
r = suma de todos los elementos **asignados** a la categoría **por el experto**
s = suma de todos los elementos **no asignados** a la categoría **por el experto**
n = suma de todos los elementos **considerados** en la evaluación

Para obtener el rendimiento general del sistema se selecciona la métrica ***F*** propuesta por Manning [1999] que combina, dentro de una misma métrica, la precisión y especificidad, sin ser afectada por el tamaño del corpus. La métrica ***F*** se calcula tomando como base la precisión *p* y la especificidad *r* (de recall) como se muestra en la fórmula siguiente:

$$F = \frac{1}{\alpha \frac{1}{p} + (1-\alpha)\frac{1}{r}}$$

Donde α representa a un factor de ponderación o ajuste en función de la importancia relativa de la precisión y especificidad para el sistema que se evalúa. Para el trabajo desarrollado se consideran de igual importancia, la precisión y la especificidad, por lo que se asigna α = 0.5 y la fórmula puede simplificarse quedando como se muestra:

$$F = \frac{2pr}{p+r}$$

Resumiendo, las métricas seleccionadas para evaluar el sistema son: *F* que nos muestra el *rendimiento*, como el mayor número de aciertos con el menor número de fallas; y la *exactitud* que nos muestra el *comportamiento* del sistema, como la capacidad de asemejarse al ser humano en la apreciación del fenómeno estudiado al "coincidir con el experto" en los aciertos y los rechazos. La precisión y especificidad se calcularán como base para el cálculo de *F* y también se reportarán o comentarán cuando se considere necesario.

## *5.3 Tamaño de la muestra*

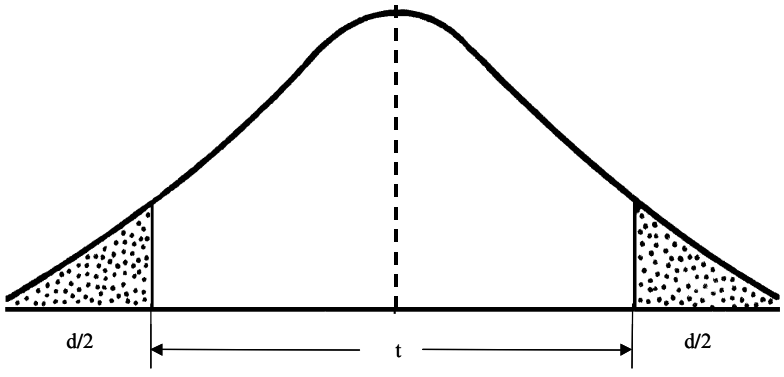
Reconociendo como base: que un parámetro poblacional correcto sólo puede obtenerse por el estudio de toda la población; que aún en este caso la certeza del valor está sujeta al proceso e instrumentos de medición; que, acorde con lo anterior, el muestreo siempre lleva involucrado

un error en la estimación del parámetro; que el muestreo es necesario en la experimentación por la imposibilidad económica, física y de tiempo; se estimó el tamaño de la muestra representativa para la evaluación del prototipo inicial.

Considerando que la coherencia textual se observa entre oraciones consecutivas y que el párrafo puede ser tomado como la unidad mínima de coherencia textual; se imprimieron los archivos juntos, como un solo archivo consecutivo, y se revisaron los primeros 100 párrafos con el objetivo de detectar la presencia de expresiones nominales (al menos una) compuestas por "det + nombre común" que presenten el fenómeno de correferencia o anáfora indirecta. La mayoría de los párrafos presentaron el fenómeno de correferencia o anáfora indirecta y se encontraron sólo 9 párrafos que contenían correferencia, pero lo llevaban a cabo usando pronombres. Con este análisis visual de los archivos se tuvo un panorama de la proporción de las muestras en la población ***p***; en otras palabras, la población de la que podrían hacerse inferencias al presentar los fenómenos de correferencia y anáfora indirecta y poder obtener resultados a partir de una muestra específica. Se puede calcular una estimación inicial de la proporción muestral ***p*** de párrafos que presenta correferencia o anáfora indirecta en un archivo a procesar como:

$$p = \frac{91}{100} = 0.91$$

Considerando un muestreo simple con distribución normal para este número de archivos el tamaño adecuado de la muestra n [Cochran et al. 1977] puede obtenerse de:




$$n = \frac{t^2 pq}{d^2}$$

Donde:

***t*** = la abscisa de la curva de la distribución normal que corta un área de riesgo en las colas. En otras palabras, es el valor de la desviación estándar, de la distribución normal, correspondiente a la probabilidad de confianza deseada.

***d*** = margen de error (varianza esperada) de la precisión *en la proporción muestral* estimada ***p***

$q = 1 - p$

Habiendo obtenido un estimado de ***p*** se puede calcular el valor de ***q*** en q = 1 - p = 0.09. Deseando mantener un margen de error menor al 10% el valor de ***d*** = 0.10, y para un nivel de confianza del 95% el valor de ***t*** = 1.96 [ver Pág. 157, Spiegel, 1976], se calcula el valor de ***n***, obteniendo que el número de muestras debe ser aproximadamente 32.

$$n = \frac{(1.96)^2(0.91)(0.09)}{(0.10)^2} = (384.16)(0.91)(0.09) = 31.46 \cong 32$$

Considerando que:

- La necesidad de obtener la información al menor costo implica terminar a tiempo el proyecto porque requerirá validación manual [Mendenhall et al. 1986].
- La distribución muestral de medias, proporciones y medianas se ajusta mucho a una normal para ***n*** igual o mayor a 30 incluso para poblaciones no normales [Spiegel 1976].
- Es el primer estudio al respecto y los resultados obtenidos pueden mejorarse junto con el prototipo en un proyecto posterior.

Se puede considerar como adecuado analizar al menos 32 archivos seleccionando sólo los del género de articulistas (ver tabla 7). En el anexo B se presenta una tabla que muestra las características y estadísticos básicos de los documentos seleccionados.

## 5.4 Resultados con el prototipo

El programa de evaluación para determinar la posibilidad de correferencia se corrió sobre los 38 documentos de articulistas dando libertad de búsqueda hacia atrás hasta siete verbos

obteniendo los resultados mostrados en la tabla 20 donde también se muestra el tiempo de procesamiento en segundos.

Los comentarios de resultados se harán sobre las corridas de programas con un documento del corpus seleccionado (a14) que tiene las características mostradas en la tabla 21; se muestra en el anexo d el ejemplo del archivo de entrada etiquetado, y el texto en formato normal de lectura en el anexo C.

| N° | Archivo | Expresiones nominales | | | Tiempo |
|---|---|---|---|---|---|
| | | Correferencia | No correferente | Total | Segundos |
| 1 | a1 | 64 | 245 | 309 | 336.03 |
| 2 | A2 | 8 | 25 | 33 | 85.47 |
| 3 | A4 | 9 | 22 | 31 | 233.76 |
| 4 | A10a | 36 | 80 | 116 | 113.53 |
| 5 | A10b | 29 | 80 | 109 | 113.42 |
| 6 | A11a | 36 | 76 | 112 | 119.85 |
| 7 | A11b | 21 | 81 | 102 | 115.78 |
| 8 | A12 | 68 | 151 | 219 | 250.85 |
| 9 | a13a | 35 | 119 | 154 | 163.45 |
| 10 | a13b | 24 | 103 | 127 | 127.38 |
| 11 | a13c | 16 | 52 | 68 | 66.9 |
| **12** | **a14** | **52** | **82** | **134** | **148.84** |
| 13 | a15a | 28 | 108 | 136 | 226.32 |
| 14 | a15b | 24 | 53 | 77 | 24.32 |
| 15 | a15c | 13 | 58 | 71 | 95.79 |
| 16 | a18 | 8 | 87 | 95 | 90.9 |
| 17 | a19 | 4 | 10 | 14 | 8.13 |
| 18 | a20 | 8 | 65 | 73 | 106.5 |
| 19 | a21a | 33 | 147 | 180 | 279.46 |
| 20 | a21b | 7 | 53 | 60 | 55.42 |
| 21 | a21c | 13 | 43 | 56 | 57.12 |
| 22 | a22a | 14 | 87 | 101 | 77.23 |
| 23 | a22b | 5 | 33 | 38 | 33.06 |
| 24 | a23a | 28 | 113 | 141 | 223.55 |
| 25 | a23b | 58 | 112 | 170 | 173.95 |
| 26 | a24 | 106 | 270 | 376 | 640.82 |
| 27 | a25a | 52 | 230 | 282 | 415.79 |
| 28 | a25b | 22 | 83 | 105 | 147.2 |

| N° | Archivo | Expresiones nominales | | | Tiempo |
|---|---|---|---|---|---|
| | | Correferencia | No correferente | Total | Segundos |
| 29 | a26a | 27 | 133 | 160 | 191.3 |
| 30 | a26b | 65 | 112 | 177 | 302.86 |
| 31 | a26c | 49 | 79 | 128 | 134.13 |
| 32 | a27 | 17 | 65 | 82 | 88.43 |
| 33 | a28a | 40 | 107 | 147 | 214.1 |
| 34 | a28b | 31 | 102 | 133 | 173.57 |
| 35 | a28c | 21 | 24 | 45 | 49.38 |
| 36 | a29 | 8 | 33 | 41 | 45.75 |
| 37 | a30a | 19 | 98 | 117 | 144.13 |
| 38 | a30b | 2 | 19 | 21 | 12.8 |
| | Suma | 1102 | 3446 | 4548 | 5887 |
| | Promedio | 29 | 91 | 120 | 154.93 |

**Tabla 20 Resultados de una corrida general**

| Descripción | Cantidad | Porcentaje |
|---|---|---|
| Adjetivos | 58 | 7.00 |
| Adverbios | 42 | 5.06 |
| Determinantes | 134 | 16.19 |
| Nombres | 200 | 24.16 |
| Verbos | 143 | 17.27 |
| Pronombres | 46 | 5.56 |
| Conjunciones | 65 | 7.85 |
| Preposiciones | 134 | 16.19 |
| Numerales | 3 | 0.36 |
| Números | 3 | 0.36 |
| **Palabras Totales** | **828** | **100.00** |

**Tabla 21 Características del documento A14**

Se utilizó un diccionario específico donde cada "entrada" de palabra está relacionada con las palabras que pueden ser sinónimos a ella. Una vez detectada o marcada la unidad léxica, obteniendo una expresión referencial, se convierte en un correferente potencial por lo que se buscan los posibles candidatos referentes anteriores, desde la oración previa hacia el inicio del texto; se determina el grado de satisfacción por similitud. Si se logra, significa que existe la relación correferencial de otra forma se supone inexistente.

El programa en una corrida libre marcó 134 nombres precedidos por un determinante. Detectó una posible correferencia (relación de sinonímia) en 52 de estos nombres con algún nombre que lo antecede en la búsqueda libre de todo el contexto lingüístico. Al verificar el número de correferencias reales en el texto (verificación manual) se encontraron sólo 21. Ante esta situación se decidió restringir la búsqueda hacia atrás (tomando en cuenta que la coherencia se da entre oraciones consecutivas) y se encontró que al restringirla a quince nombres se detectaban sólo 24 con posible relación de sinonímia dentro de los cuales se encontraban los 21 correferentes verificados manualmente, en esta situación se alcanza una precisión del 87.50% con una exactitud del 88.88% ; los resultados se concentran en la tabla 22.

| Evaluación | Total | Correferentes | No-corref | Real | Precisión | Exactitud |
|---|---|---|---|---|---|---|
| Libre | **134** | **52** | **82** | **21** | 40.38 % | 76.87 % |
| Restringida | 27 | 24 | 3 | 21 | 87.50 % | 88.88 % |

**Tabla 22 Resumen de resultados en a14 con diccionario específico**

| | | **EXPERTO** | | |
|---|---|---|---|---|
| | | **Si** | **No** | |
| **SISTEMA** | **Si** | 21 | 31 | **52** |
| | **No** | 0 | 82 | **82** |
| | | **21** | 113 | **134** |

**Tabla 23 Ejemplo de cálculo de métricas**

En la tabla 23 se substituyen los valores obtenidos en la corrida, marcados con **negrita**, para mostrar un ejemplo del cálculo de las métricas del primer renglón de la tabla 22.

$$precisión = \frac{21}{52} = .4038 \qquad especificidad = \frac{21}{21} = 1 \qquad exactitud = \frac{21+82}{134} = .7687$$

Relacionando la información, del primer renglón de la tabla 22 con las características del documento en la tabla 21, se puede observar que existen 46 pronombres que "normalmente" contienen correferencias por medio del fenómeno de anáfora directa con lo cual esperaríamos 134 – (46+21) = 67 posibilidades de: anáforas indirectas o referencias que sean parte de la información complementaria del documento. La especificidad es alta (100%) porque se utilizó un diccionario de sinónimos construido específicamente para este documento; en este caso la precisión y la exactitud aumentan al restringir la búsqueda hacia atrás.

Estos resultados animaron la implantación del algoritmo de resolución de anáfora indirecta, para trabajar con nombres comunes; con el programa se hizo una corrida libre y una corrida restringida a diez verbos, obteniendo los resultados que se muestran en la tabla 24. Cabe mencionar que en la verificación manual se encontraron 23 casos de anáfora indirecta en el texto por lo que la precisión y la exactitud de la anáfora indirecta se calculan con respecto a este concepto.

| | | Programa | | | Real | | Ana Ind en % | |
|---|---|---|---|---|---|---|---|---|
| Evaluación | Total | Corr | AInd | No-corr | Corr | AInd | Precisión | Exactitud |
| Libre | 134 | 65 | 27 | 42 | 21 | 23 | 85.19 | 97.01 |
| Restringida | 134 | 25 | 25 | 83 | 21 | 23 | 92.00 | 98.51 |

**Tabla 24 Resultados en a14 con anáfora indirecta**

La especificidad es alta (100%) porque se utilizó un diccionario de sinónimos construido específicamente para este documento y lo mismo puede decirse de la precisión y la exactitud. La validez de estas pruebas radica en probar que el modelo y el algoritmo son adecuados, aunque altamente dependientes de la información apropiada en el diccionario de escenarios.

Buscando que el sistema pueda trabajar con otro archivo diferente a la muestra seleccionada, se repitió el experimento para la evaluación de correferencias con el mismo archivo (a14) utilizando el diccionario de sinónimos del Laboratorio de Lenguaje Natural del CIC-IPN capturado por medio de un escáner, esperando obtener resultados iguales o muy parecidos. El programa marcó 134 nombres precedidos por un determinante. Detectó una posible relación de sinonímia en 73 de estos nombres con algún nombre que lo antecede en la búsqueda libre de todo el contexto lingüístico. Al tener validadas sólo 21 correferencias reales, se decidió restringir la búsqueda hacia atrás y se encontró que al restringir el inicio de búsqueda a la oración previa y hasta cuatro nombres anteriores se detectaban sólo 29 con posible relación de sinonímia dentro de los cuales se encontraban los 21 correferentes verificados manualmente, en esta situación se alcanza una precisión del 72.42%; los resultados se concentran en la tabla 25.

La diferencia de resultados, disminución drástica de la precisión (22 y 15 puntos porcentuales), al cambiar el diccionario de sinónimos ha obligado a revisar el diccionario de sinónimos del Laboratorio de Lenguaje Natural encontrando errores debidos al proceso de

captura por medio de un escáner, su proceso de corrección fue descrito como preparación del diccionario de sinónimos (en la sección 4.4).

| Corrida | Total | Correferentes | No-corref | real | Precisión | Exactitud |
|---|---|---|---|---|---|---|
| Libre | 134 | 73 | 61 | 21 | 28.77 % | 61.19 % |
| Restringido | 110 | 29 | 81 | 21 | 72.42 % | 92.73 % |

**Tabla 25 Resumen de resultados con diccionario del LLN CIC-IPN**

Después de corregir el diccionario de sinónimos y obtener un diccionario de escenarios, descrito en la sección 4.5, la atención se concentró en las pruebas para determinar un tamaño de la ventana de búsqueda hacia atrás para poder trabajar con texto libre.

## *5.5 Tamaño de ventana de búsqueda*

En la sección anterior se mencionaron corridas libres en la búsqueda de la correferencia o anáfora indirecta potencial "desde la posición actual hasta el inicio del archivo" y corridas restringidas en función del número de ocurrencias de una bandera definida por ejemplo: 7 verbos, 15 nombres, 10 verbos, 4 nombres. Estas corridas tenían como intención mantener el contexto lingüístico en memoria para lograr que los resultados incluyeran todas correferencias y anáforas indirectas, detectadas en la verificación manual, porque los diccionarios, de sinónimos y escenarios, contenían la información completa.

La intención original "*incluir el contexto lingüístico (los antecedentes necesarios) que satisfaga las correferencias y anáforas indirectas*" sigue siendo válida y esto puede lograrse de dos formas:

- almacenar todas las unidades léxicas, estructura e información implícita (de sinónimos y escenarios) y mantener el sistema dentro del límite físico de 4800 unidades léxicas (o tokens) o 45 KB aproximadamente del tamaño de archivo a procesar
- modelar al lector humano que almacena en el contexto lingüístico sólo la información relevante (nombres propios; enlaces correferenciales y de anáforas indirectas; y los últimos N grupos de unidades léxicas necesarios para satisfacer las correferencias y anáforas indirectas.

La primera alternativa fue apropiada para el desarrollo del prototipo inicial porque permitió hacer corridas con el tamaño necesario de la ventana; pero para poder alcanzar la segunda meta o alternativa se plantean dos problemas: ¿cuál es la bandera o marcador más adecuado? y ¿cuál es el tamaño N apropiado de la ventana de búsqueda?

Para plantear mejor el problema de determinar la bandera adecuada se utilizará parte del texto del archivo a14 con "oraciones" numeradas (considerando como oración la separación de puntuación conocida como "punto"). Se han marcado con **negrita** los verbos y con *cursiva* los nombres de la expresión nominal.

1. Cuando **escribo** esto la *Madre_Coraje* peruana **acaba** de **ser** reventada por los *senderistas*.
2. **Veo** su *foto* en los *periódicos*: una *mujer* joven, atractiva, probablemente zamba, esto_**es**, mestiza de negra e india; oscura de color, en_fin, como **son** oscuros todos los *habitantes* de las *villas* limeñas, *arrabales* de miseria en donde se **hacinan** cientos de miles de *personas*.
3. **Son**, en su **mayoría**, **indígenas** que **bajaron** de los *Andes* **huyendo** del *hambre*, del *atraso* y la *tuberculosis*; **quisieron llegar** a la *ciudad*, pero **quedaron varados** en las *afueras*, a una decena de *kilómetros*, en los sórdidos *arenales* que **rodean** *Lima*, en donde **plantaron** sus *chabolas*, precarios *tenderetes* de cartón y *cajones* astillados.
4. El *liderazgo* de *María_Elena* **nació** de aquella *miseria* y de una increíble *voluntad* de superación.
5. De la *generosidad*, de la *inteligencia*, del *tesón*.

Una revisión general de las cinco oraciones permite observar el diferente tamaño y composición, como se muestra en el resumen de la tabla 26.

| Oración | Verbos | Nombres | Puntuación | palabras |
|---|---|---|---|---|
| 1 | 3 | 2 | 1 | 14 |
| 2 | 4 | 7 | 10 | 45 |
| 3 | 9 | 14 | 10 | 53 |
| 4 | 1 | 4 | 1 | 16 |
| 5 | 0 | 3 | 3 | 8 |

**Tabla 26 Resumen de elementos de oraciones**

El sólo signo de puntuación "punto" ("punto y seguido" o "punto y aparte") contiene diferente número de nombres de acuerdo a la extensión de las expresiones nominales en la oración; además de la posibilidad de confusión con el punto que acompaña a las abreviaturas (Dr., Sr., etc.). Se pueden reconocer contrastes entre la oración 1 y 4; la oración 4 con sólo un verbo agrupa hasta 4 nombres mientras la oración 1 con 3 verbos sólo agrupa a 2 nombres. Las

oraciones 2 y 3 tienen el mismo número de signos de puntuación y casi el mismo número de palabras pero la oración 3 tiene mayor número de verbos y nombres. La oración 5 no tiene verbo explícito, debido al fenómeno de elipsis verbal, sin embargo contiene hasta 3 nombres y 3 signos de puntuación. Estas observaciones presentan un panorama confuso para determinar el marcador adecuado.

Con el fin de observar el comportamiento de cada tipo de bandera se decidió hacer corridas de diferente tamaño para detectar la correferencia utilizando uno de los archivos de texto libre, "Contra la guerra" (ver tabla 28), obteniendo los resultados que se grafican de la figura 19 a la figura 22; los resultados completos se pueden apreciar tabulados en el anexo G.

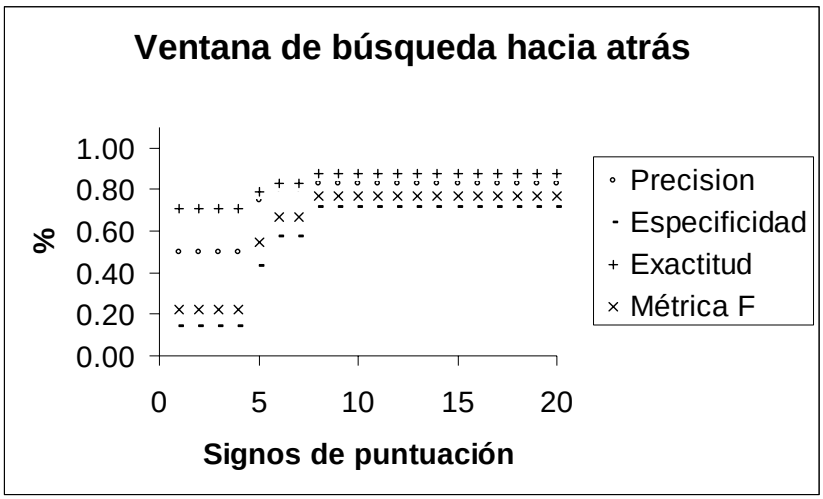



**Figura 19 Evaluación de signos de puntuación como bandera**

En todas las figuras se puede observar que conforme aumenta el tamaño de la ventana (número de signos de puntuación, nombres, verbos o el "punto") mejoran los valores de las métricas utilizadas porque aumenta el número de elementos detectados, hasta alcanzar un valor constante una vez que han sido incluidas todas las correferencias que pueden ser detectadas; después de este punto ya no se incrementan los valores a pesar del aumento de tamaño de la ventana.

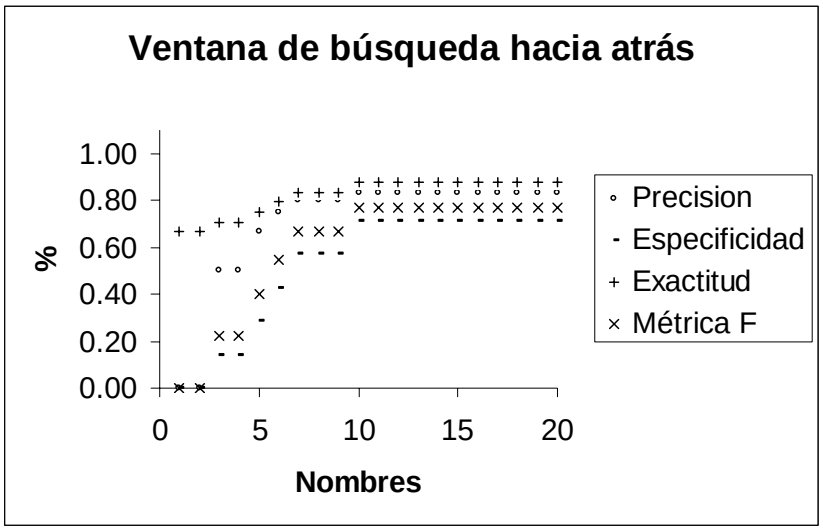



**Figura 20 Evaluación de los nombres como bandera**

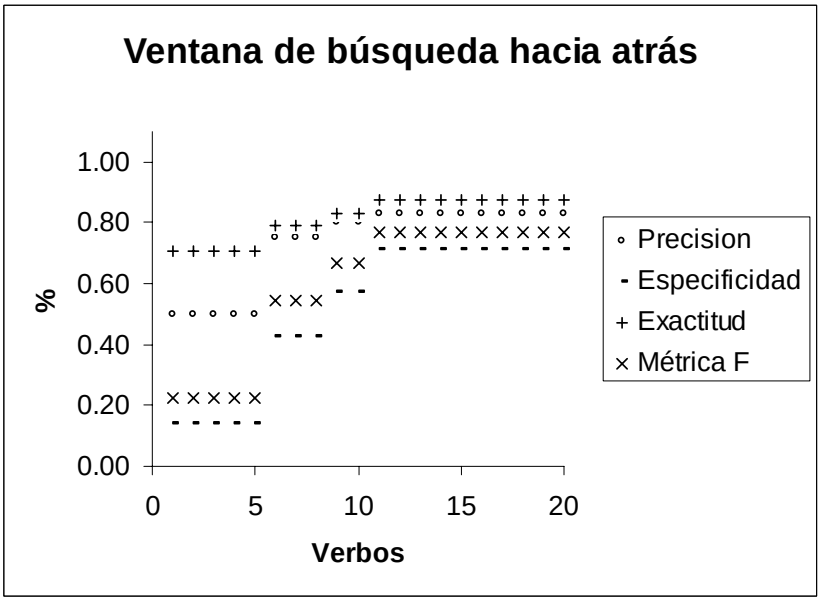



**Figura 21 Evaluación como bandera de los verbos**

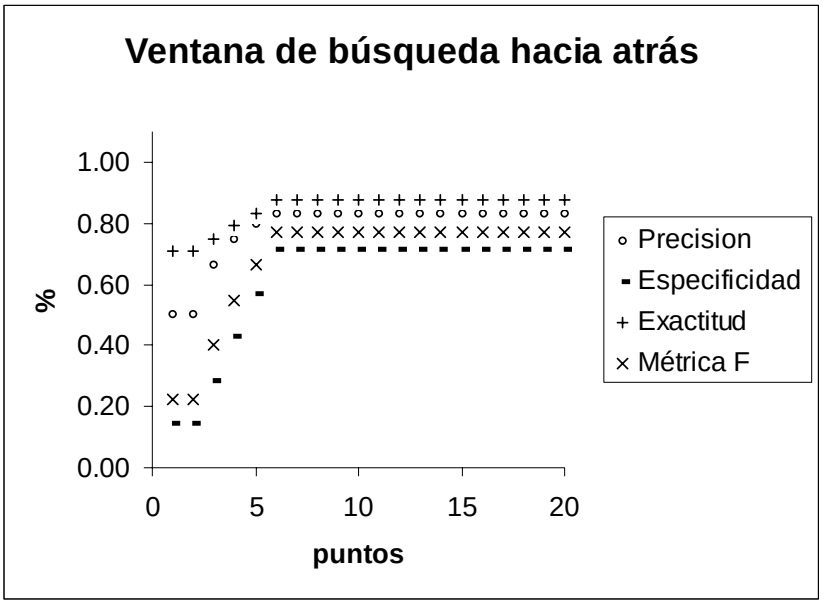



**Figura 22 Evaluación como bandera de los puntos**

En otras palabras, una ventana de tamaño mayor a este límite, de N banderas encontradas, incrementa sin necesitarlo el tiempo de procesamiento y la memoria requerida para mantener el contexto lingüístico; una ventana menor a este límite afectará adversamente a la precisión, especificidad y exactitud obtenidas en la corrida (por fallas en la detección).

Con esta información ya se puede responder a las preguntas: ¿cuál es la bandera o marcador más adecuado? y ¿cuál es el tamaño N apropiado de la ventana de búsqueda?

De acuerdo a los resultados obtenidos cualquier tipo de bandera (signo de puntuación, punto, verbo o nombre) cumple los objetivos de la búsqueda hacia atrás ya que lo más importante es el tamaño de ventana apropiado. Considerando que el punto separa las oraciones cuyo sentido está completo o que representan ideas completas expresadas en el texto se selecciona este marcador como bandera.

El tamaño apropiado de N sería **seis**, de acuerdo al resultado obtenido y representado en la figura 22 , pero hay que tomar en cuenta dos cosas: la recomendación de sicolingüístas que, conforme a los resultados de sus experimentos, recomiendan al menos **siete** como la capacidad del procesamiento de información del ser humano, lo que influye en la redacción y lectura de textos [Miller, 1956]; además es necesario considerar el riesgo de que existan abreviaturas, que

afecten al etiquetador y repercuta en los resultados de la evaluación; por lo tanto, es necesario aumentarlo a un valor que asegure "*incluir el contexto lingüístico suficiente para satisfacer las correferencias y anáforas indirectas presentes en el texto"*; así, el tamaño apropiado elegido es **nueve** como primera aproximación.

Resumiendo, la bandera más adecuada es el "punto" y el tamaño apropiado de la ventana es **nueve**. Lo anterior, concuerda con el conocimiento lingüístico de que "la coherencia textual se presenta entre oraciones consecutivas"; el "punto" agrupa ideas completas independientemente de la complejidad de la oración; permite modelar un "contexto lingüístico" emulando al lector humano; e implementarlo, sin exceder la capacidad de memoria, disponible en las computadoras actuales.

En la tabla 27 se presentan los tiempos registrados para el tipo de bandera "punto" y diferentes tamaños de ventana.

| **Archivo El Cerebro bandera "Punto"** | | | |
|---|---|---|---|
| **Tamaño Ventana** | **Corrida** | | **Formato** |
| | **Hora final** | **Duración (Seg)** | **mm:ss.dd** |
| 2 | 6:49:01 | 00:03.4 | 49:01.1 |
| 3 | 6:48:58 | 00:05.0 | 48:57.7 |
| 4 | 6:48:53 | 00:06.7 | 48:52.7 |
| 5 | 6:48:46 | 00:08.0 | 48:46.0 |
| 6 | 6:48:38 | 00:09.7 | 48:38.0 |
| 7 | 6:48:28 | 00:11.1 | 48:28.3 |
| 8 | 6:48:17 | 00:12.4 | 48:17.2 |
| **9** | **6:48:05** | **00:13.5** | **48:04.8** |
| 10 | 6:47:51 | 00:14.1 | 47:51.3 |
| 11 | 6:47:37 | 00:14.5 | 47:37.3 |
| 12 | 6:47:23 | 00:15.4 | 47:22.8 |
| 13 | 6:47:07 | 00:15.8 | 47:07.4 |
| 14 | 6:46:52 | 00:16.2 | 46:51.7 |
| 15 | 6:46:35 | 00:16.7 | 46:35.4 |
| 16 | 6:46:19 | 00:16.9 | 46:18.7 |
| 17 | 6:46:02 | 00:17.0 | 46:01.8 |
| 18 | 6:45:45 | 00:17.1 | 45:44.8 |
| 19 | 6:45:28 | 00:17.3 | 45:27.6 |
| 20 | 6:45:10 | 00:17.4 | 45:10.4 |

**Tabla 27 Duración de corrida para diferentes tamaños de ventana**

Se puede apreciar la variación directa del tiempo con el cambio del tamaño de ventana, sin embargo la principal ganancia fue la reducción de tiempo lograda con la reducción del tamaño de almacenamiento de los diccionarios al simplificar la estructura, evitar repetición de entradas y duplicidad de información.

Estos valores fueron obtenidos en una computadora LapTop Pentium IV, a 1.2G y 256 MB en RAM en un archivo de 583 palabras de texto libre incluyendo todo el proceso de conversión de texto libre a archivo sin etiquetas, etiquetado de archivo y evaluación de anáfora indirecta.

## 5.6 Resultados con archivos de texto libre

Para las pruebas de texto libre se tomaron tres archivos recibidos por e-mail en febrero del 2003; dos son comentarios contra la guerra; el tercero es una reflexión sobre el cerebro y cuatro archivos del corpus CliC-TALP previamente convertidos a texto plano con las características presentadas en la tabla 28.

Los resultados completos de las evaluaciones se presentan en el anexo J y un resumen de resultados, con el tamaño de ventana igual a nueve, se muestra en la tabla 29, donde las abreviaturas en las columnas significan: **V** = verdadero, **F** = falso o error de identificación, **Ob**=total obtenido por el programa (verdaderos + falsos), **tot det** = total de nombres precedidos por determinativos o expresiones referenciales que son candidatos para ser considerados anáfora directa, indirecta con o sin correferencia y referencia (nueva).

| Documento | Palabras | Párrafos | Líneas | KB |
|---|---|---|---|---|
| A14 | 827 | 40 | 86 | 6 |
| A12a | 574 | 19 | 51 | 4 |
| A12b | 547 | 20 | 51 | 4 |
| A12c | 271 | 9 | 22 | 2 |
| Contra la guerra | 208 | 17 | 23 | 2 |
| Instituto Oriente | 327 | 20 | 44 | 3 |
| El cerebro | 583 | 26 | 64 | 4 |

**Tabla 28 Características de Archivos para prueba de texto libre**

| | Anáfora | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Con correferencia | | | | | | Sin Correferencia | | | Total | | | |
| | Directa | | | Indirecta | | | Indirecta | | | Indirecta | | | Det |
| Archivo | V | F | Ob | V | F | Ob | V | F | Ob | V | F | Ob | |
| A12a | 4 | 4 | 8 | 9 | 15 | 24 | 3 | 8 | 11 | 12 | 23 | 35 | 67 |
| A12b | 14 | 1 | 15 | 14 | 14 | 28 | 2 | 10 | 12 | 16 | 24 | 40 | 77 |
| A12c | 11 | 0 | 11 | 5 | 1 | 6 | 3 | 4 | 7 | 8 | 5 | 13 | 39 |
| A14 | 21 | 1 | 22 | 13 | 20 | 33 | 4 | 25 | 29 | 17 | 45 | 62 | 118 |
| Cerebro | 15 | 1 | 16 | 5 | 12 | 17 | 3 | 16 | 19 | 8 | 28 | 36 | 66 |
| Contra | 7 | 1 | 8 | 0 | 5 | 5 | 0 | 2 | 2 | 0 | 7 | 7 | 24 |
| Io | 18 | 0 | 18 | 1 | 9 | 10 | 3 | 3 | 6 | 4 | 12 | 16 | 56 |

**Tabla 29 Resultados del programa para prueba de texto libre**

En la tabla 30 se presentan los valores reales identificados gracias al apoyo del M. en C. César A. Aguilar quien es un lingüista especialista en anáfora, del Grupo de Ingeniería Lingüística de la UNAM. Con la identificación desarrollada se pudieron validar los resultados del programa y desarrollar los cálculos apropiados de acuerdo a las métricas seleccionadas para la evaluación.

| Archivo | Con correferencia | | Sin Correferencia Indirecta | Total Indirecta | Det |
|---|---|---|---|---|---|
| | Directa | Indirecta | | | |
| A12a | 4 | 9 | 7 | 16 | 67 |
| A12b | 17 | 14 | 2 | 16 | 77 |
| A12c | 11 | 5 | 5 | 10 | 39 |
| A14 | 26 | 13 | 6 | 19 | 118 |
| Cerebro | 17 | 5 | 5 | 10 | 66 |
| Contra | 7 | 0 | 0 | 0 | 24 |
| Io | 23 | 1 | 5 | 6 | 56 |

**Tabla 30 Resultados reales en verificación manual de anáfora**

Se implementó el método propuesto por Gelbukh y Sidorov [1999], como una primera aproximación y sin los algoritmos de ponderación, para tener una línea base de comparación con el método desarrollado. En la tabla 31 se presentan: los resultados obtenidos con la implementación del método de Gelbukh y Sidorov (marcados con asterisco *) junto con los resultados del método desarrollado; los valores *F*, precisión, especificidad y exactitud obtenidos para cada archivo con un tamaño de ventana igual a nueve; y el promedio de todas las evaluaciones excepto las del archivo “contra” que no contiene el fenómeno de anáfora indirecta como puede verificarse en la tabla 30.

| Archivo | Métrica | | | |
|---|---|---|---|---|
| | F | Precisión | Especificidad | Exactitud |
| *A12a | 0.26 | 0.20 | 0.38 | 0.36 |
| A12a | 0.47 | 0.34 | 0.75 | 0.60 |
| *A12b | 0.12 | 0.09 | 0.19 | 0.18 |
| A12b | 0.57 | 0.40 | 1.00 | 0.69 |
| *A12c | 0.15 | 0.20 | 0.10 | 0.18 |
| A12c | 0.70 | 0.62 | 0.80 | 0.82 |
| *A14 | 0.17 | 0.53 | 0.25 | 0.41 |
| A14 | 0.42 | 0.27 | 0.89 | 0.60 |
| *Cerebro | 0.17 | 0.12 | 0.30 | 0.14 |
| Cerebro | 0.35 | 0.22 | 0.80 | 0.55 |
| *Io | 0.26 | 0.18 | 0.50 | 0.58 |
| Io | 0.36 | 0.25 | 0.67 | 0.75 |
| *Promedio_base | 0.21 | 0.15 | 0.35 | 0.27 |
| Promedio | 0.48 | 0.35 | 0.82 | 0.67 |
| **Diferencia** | **0.27** | **0.20** | **0.47** | **0.40** |

**Tabla 31 Evaluación inicial de la anáfora indirecta**

Como puede observarse, se logró obtener una mejora (ver renglón Diferencia en tabla 31), comparado el método de Gelbukh y Sidorov, pero aún no se alcanzaban los obtenidos previamente en el prototipo inicial; para obtener una explicación se hicieron evaluaciones variando el tamaño de ventana para cada fenómeno de forma independiente: correferencia directa, correferencia indirecta y anáfora indirecta no correferencial; además se verificó en los archivos de seguimiento las razones que permitieran mejorar los resultados.

Esta búsqueda llevó a determinar las causas principales que afectaban el rendimiento del sistema encontrando:

I. Errores en el proceso de etiquetado de TnT

II. Expresiones referenciales marcadas con relaciones predefinidas

III. Falta de información o información incorrecta en los diccionarios de sinónimos y de escenarios

IV. La interrelación entre los fenómenos se ve afectada por el problema de polisemia que provoca ambigüedad.

**I** Se solucionaron los errores en el proceso de etiquetado de TnT corrigiendo manualmente los archivos después del proceso de etiquetado para simular la utilización de un etiquetador "perfecto" (desarrollar uno nuevo queda fuera del alcance de esta tesis).

**II** Para corregir el efecto de las expresiones referenciales predefinidas se modificó el programa para marcarlas antes de efectuar la evaluación de la anáfora y no tomarlas en cuenta en el proceso de evaluación. Una expresión referencial predefinida se presenta en la oración debido al uso de la preposición "de" que significa propiedad, posesión, materia, origen, etc. por ejemplo:

(88) Es admirable *la inteligencia* **de** Ismael

(89) Dame *esa hoja* **de** papel

(90) Me regalaron *este crucifijo* **de** madera

En los ejemplos (88) al (90) se puede observar que las expresiones referenciales (det + nombre_común + **de** + nombre) no deben evaluarse como candidatos de anáfora porque su relación en el texto ya está predefinida con respecto al nombre que le sigue después de la preposición "de".

**III** Se corrigió la información incorrecta en el diccionario de sinónimos gracias al apoyo de la Lic. Martha Grizel Delgado Rodríguez quien es una lingüista, del Grupo de Ingeniería Lingüística de la UNAM, y que manualmente validó la información en el diccionario. Se solucionó manualmente la falta de información en el diccionario de sinónimos por el autor de esta tesis suministrando la información faltante de acuerdo al diccionario de sinónimos de Manuel Seco de editorial ESPASA-CALPE. Se corrigió la información de WordNet en Español de acuerdo al contexto Mexicano como se reportó en la preparación del diccionario de escenarios (ver sección 4.5).

**IV** El problema de polisemia provocaba que una misma forma de palabra o nombre común se marque como correferencia, directa o indirecta, o como anáfora indirecta sin serlo y por el orden en que se evalúan afecta la resolución de el paso siguiente; esto hace que la relación se encuentre en el diccionario pero que en el texto no sea aplicable para ese caso específico. Se solucionó parcialmente haciendo evaluaciones con diferentes tamaños de ventana para cada

fenómeno de forma independiente para determinar el tamaño de ventana apropiado para cada fenómeno en cada archivo.

El análisis de los resultados permitió observar que el tamaño de ventana con valor de **nueve** seleccionado permitía encontrar la mayoría de las relaciones pero también provocaba interferencia en la evaluación de los fenómenos en conjunto. Utilizando el valor de *F* como indicador del mejor rendimiento global en cada caso analizado se encontró, buscando manualmente en las tablas mostradas en el anexo J, que cada fenómeno requería diferente tamaño de ventana para cada archivo y el resumen de estas observaciones se presenta en la tabla 32. En esta tabla puede observarse que para el caso de la correferencia directa el tamaño de ventana es más amplio, abarcando todo el documento en algunos archivos; para el caso de anáfora indirecta (con y sin correferencia) el tamaño de ventana queda comprendido o es inferior al **nueve** seleccionado corroborando las observaciones hechas en el análisis del tamaño de ventana (ver sección 5.5).

| **Archivo** | **con correferencia** | | | | **sin correferencia** | |
|---|---|---|---|---|---|---|
| | **Directa** | | **Indirecta** | | **Indirecta** | |
| | **Vent** | ***F*** | **Vent** | ***F*** | **Vent** | ***F*** |
| **A12a** | 3 | 0.89 | 2 | 0.67 | 2 | 0.63 |
| **A12b** | 10 | 0.97 | 4 | 0.70 | 2 | 0.40 |
| **A12c** | 9 | 1.00 | **5** | **0.83** | 4 | 0.55 |
| **A14** | 20 | 0.98 | 3 | 0.69 | 2 | 0.44 |
| **cerebro** | 14 | 0.92 | 4 | 0.50 | **1** | **0.80** |
| **Io** | 18 | 1.00 | 1 | 0.67 | 5 | 0.57 |

**Tabla 32 Mejor ventana de acuerdo al valor de *F***

De acuerdo a lo anterior, se modificó el programa para poder utilizar una ventana diferente en cada fenómeno y poder simular la detección y resolución de la anáfora indirecta "suponiendo" que se cuenta con mejores métodos para resolver la correferencia obteniendo los resultados que se muestran en la tabla 33.

Con esta adecuación se logró una mejora substancial comparada con los resultados presentados en la tabla 31: de 12 puntos en la métrica *F*, 16 puntos en la precisión y 15 puntos en la exactitud a pesar de una reducción de 8 puntos en la especificidad. En otras palabras, se mejoró

notablemente el rendimiento, hasta alcanzar un valor de *F* igual a 0.60, y el comportamiento del sistema, hasta alcanzar una exactitud en el acuerdo con el "experto humano" de 0.82.

| Archivo | Métrica | | | |
|---|---|---|---|---|
| | F | Precisión | Especificidad | Exactitud |
| A12a | 0.65 | 0.54 | 0.81 | 0.79 |
| A12b | 0.60 | 0.45 | 0.88 | 0.75 |
| A12c | 0.73 | 0.67 | 0.80 | 0.85 |
| A14 | 0.58 | 0.48 | 0.74 | 0.83 |
| Cerebro | 0.64 | 0.58 | 0.70 | 0.88 |
| Io | 0.40 | 0.33 | 0.50 | 0.84 |
| **Promedio** | **0.60** | **0.51** | **0.74** | **0.82** |
| *Promedio_base | 0.21 | 0.15 | 0.35 | 0.27 |
| **Diferencia** | **0.39** | **0.36** | **0.39** | **0.55** |

**Tabla 33 Evaluación final de la anáfora indirecta**

Si los resultados se comparan con los obtenidos con la implementación del método de Gelbukh y Sidorov (marcado como *promedio_base en la tabla 33) se alcanza una mejora: de 39 puntos en la métrica F, 36 puntos en la precisión, 39 puntos en la especificidad y 55 puntos en la exactitud.

A continuación se presenta la salida del programa ante la presencia de correferencia directa, correferencia indirecta y anáfora indirecta donde se marcan los errores detectados.

Archivo f_tnt\contra.tts
(42) * El problema >ci-->(17) guerra . ERROR: del modelo
( 52) * guerra ←cd-(66) guerra
( 78) * mundo ←ai-(108) pueblo
( 91) * paz ←ci-(236) abrazo
( 97) * favor ←ci-(135) nombre ERROR: del modelo
(103) * gracias ←ci-(156) derechos
(127) * malos ←ai-(221) vecino ERROR: del modelo
(151) * el mal >ci--> (42) problema es ERROR: del modelo
(169) * partes ←ai-(196) lado
(232) * propio ojo >cd--> (219) ojo del ERROR: no correferente

En los casos de "ERROR: del modelo" la información se encuentra en el diccionario y la relación existe; además la ventana de búsqueda hacia atrás incluye esta palabra; dando una discrepancia con el lector humano. En el caso de "ERROR: no correferente" se utiliza la misma

palabra y concepto pero se refieren a objetos diferentes del mundo real. Estos son los tipos de errores que quedan pendientes de resolver en trabajo futuro.

A continuación se muestra una parte del texto extraído de la salida del programa en formato htm. Se han marcado con subrayado y **negrita** las palabras involucradas en la situación que provoca el error.

En todo el mundo (8)10000000 , dicen , hemos desfilado contra la **guerra** (17).

Participé durante un trecho (22)en la de Barcelona (26).

Era impresionante .

Dicen que éramos 1300000 personas (35)clamando por la paz (39).

El **problema** (42)es por qué no sucede lo mismo ante cualquier guerra (52)o pisoteo (54)de los derechos (57)humanos .

No se trata sólo de esta guerra (66), sino de muchas que se están librando por todo el mundo (78).

Además creo que debería quedar claro que el marchar por la paz (91)no significa que estemos a **favor** (97)de Sadam (99), pues " gracias (103)" a él su pueblo (108)se está hundiendo .

¿ Vivimos realmente en un mundo (118)de buenos (120)y malos (122)? ¿ Están los **malos**(127) legitimados para permitir que , en su **nombre** (135), mueran inocentes (138)? ¿ Está tan claro lo que es el bien (148)y el **mal** (151)? Hemos elaborado los derechos (156)humanos y parece que no sirven para nada , pues en todas partes (169)se siguen atropellando .

Quizá llegue el momento (177)en que muchos millones nos manifestemos en contra de este atropello (188), que a menudo sucede a nuestro lado (196).

No hace falta alejarse mucho para observarlo .

Como dice la sabiduría (209)popular " es fácil ver la paja (216)en el **ojo** (219)del **vecino** (221), sin darse cuenta (225)de la viga (228)en nuestro propio **ojo** (232)" .